Eltiraĵo el
Acta Sanmarinensia I.L 3/1990
ISBN 3-88064-180-3

Prezentata en TTT sub la URL:
http://www.forst.tu-muenchen.de/publ/quednau/baza.html

BAZA STATISTIKA METODARO

de H. D. Quednau, München (D)

(Kurso prezentita dum SUS 6 en San Marino, aŭgusto/septembro 1989)

Resumo

La kurso pritraktas la bazajn konceptojn de la **aplikata** statistiko. Ĝi direktas sin al la uzantoj de statistikaj proceduroj kaj tial ne priparolas metodojn kaj rezultojn de la **matematika** statistiko. Ne estas postulataj antaŭkonoj superantaj la komprenon de la simpla matematika simbolaro.

Estas pritraktataj la bazaj ecoj de statistikaj datenaroj kaj kelkaj statistikaj adedoj, kiel aritmo, varianco, pluriloj, korelaci- kaj regresi-koeficientoj. Kadre de la konkluda statistiko priparolatas la dunomiala kaj la gaŭsa distribuoj, kiel ekzemploj de kontinua kaj malkontinua distribuoj. Aparte granda graveco estas atribuata al la kompreno de la esenco de statistikaj testoj. Ili estas ekzempligitaj per testo pri donita p_0 de dunomiala distribuo kaj de 3 aplikoj de la t-testo (testo pri donita μ_0, testo pri malsamaj ekspektoj ĉe du ne-interdependaj samploj kaj testo pri donita deklivo kadre de la simpla lineara regresi-analizo).

Zusammenfassung

Der Kurs behandelt die grundlegenden Konzepte der **angewandten** Statistik. Er richtet sich an den Anwender statistischer Verfahren und geht daher nicht auf Methoden und Ergebnisse der **mathematischen** Statistik ein. Es werden keine Kenntnisse vorausgesetzt, die über das Verständnis der einfachen mathematischen Symbolik hinausgehen.

Es werden die grundlegenden Eigenschaften statistischer Datensätze und einige statistische Maßzahlen, wie Mittelwert, Varianz, Quantile sowie Korrelations- und Regressionskoeffizienten besprochen. Im Rahmen der schließenden Statistik werden Binomial- und Normalverteilung als Beispiele für eine diskrete und eine kontinuierliche Verteilung behandelt. Besonderer Wert wird auf das Verständnis des Wesens statistischer Tests gelegt, die am Beispiel des Tests auf vorgegebenes p_0 der Binomialverteilung und von 3 Anwendungen des t-Tests (Test auf vorgegebenes μ_0 , Test auf

unterschiedliche Erwartungswerte bei 2 unabhängigen Stichproben und Test auf vorgegebene Steigung bei der einfachen linearen Regressionsanalyse) besprochen werden.

I. Enkonduko

Tiu kurso prezentas la plej bazajn metodojn de la **aplikata statistiko**, kiuj iĝas pli kaj pli ne-malhaveblaj por ĉiu persono, kiu okupiĝas pri esploroj en la natur- kaj soci-sciencoj. La kurso **ne** pritraktas la matematikan statistikon, sed ĝi intencas komprenigi ankaŭ al nematematikistoj la bazan statistikan metodaron, por ke ili povu ĝin korekte apliki en siaj propraj sciencoj. La prezentado de matematiko estas limigata al la nepra minimumo.

Unu el la plej bazaj ŝtupoj por la akiro de sciencaj ekkonoj estas tre ofte (kvankam ne ĉiam !) la kolektado de datenoj. Ofte tiuj datenoj enhavas hazardajn fluktuaĵojn kiuj malfaciligas ilian interpretadon. Oni pensu ekzemple pri datenoj el demoskopiaj enketiloj, el biologiaj esplorprotokoloj aŭ el medicinaj ekzamenoj. Estas la tasko de la aplikata statistiko provizi rimedojn al la esploranto por akiri science funditajn asertojn el tiaj datenaroj. Ĝi montras unue, kiamaniere oni organizu esploron kaj due, kiamaniere oni utiligu la akiritajn datenojn por ricevi plejeble multajn informojn. Ĉi-tie ni okupiĝos nur pri la dua ŝtupo, do la prilaborado de jam akiritaj datenoj, ĉar la statistike fundita esplor-planado estas ankoraŭ tro komplika por enkonduka kurso.

II. Statistikaj datenaroj

La esploro, kiu liveras al ni la prilaborotan datenaron, povas esti aŭ **enketo** aŭ **eksperimento**. La diferenco inter tiuj estas, ke dum eksperimento ĉiuj efikoj kiuj eble povus influi la esploratan celvariablon, estas fiksitaj de la eksperimentanto. Eksperimentoj estas la tipaj esploroj en la natursciencoj. Kontraŭe, dum enketo la esploranto ne povas fiksi la influantajn efikojn, li nur povas plejeble komplete protokoli ilin. Enketoj estas tipaj en la soci-sciencoj. Sed ankaŭ en epidemiologiaj kaj ekologiaj esploroj, ekzemple en la forstscienco, ili estas ofte neeviteblaj.

Ekzemplo : Oni volas esplori la efikon de SO_2 al la kresko de piceoj. Tion oni povas fari unue per **eksperimento** (taŭga por arbidoj) : Oni traktas piceidojn de difinitaj genotipo kaj aĝo, kreskigitajn en difinitaj kondiĉoj (temperaturo,grundo,aermalseko) en fitotrono* per antaŭfiksitaj kvantoj de SO_2 kaj mezuras la jaran kreskon. Dua eblo estas esploro per **enketo** (taŭga por nejunaj arboj) : Oni metas kradon super la esplorendan areon kaj protokolas por ĉiu piceo, kiu situas plej proksime al la opa krad-punkto, krom la lastjara kresko kaj la SO_2-koncentriteco ankaŭ kelkajn pluajn mezuraĵojn, kiuj koncernas aŭ la arbon mem (alto, diametro) aŭ la lokon (alteco super mar-nivelo, ecoj de la grundo).

La datenaron kiun liveras la esploro (ĉu enketo, ĉu eksperimento) oni kutime aranĝas en formo de longa matrico, la **esplor-matrico**. La horizontaloj de la matrico enhavas la datenojn kiuj apartenas al la sama **esplor-unuo**. La esplor-unuo povas esti arbo, besto, homo, sed ankaŭ folio, peco de histo aŭ tuta hospitalo aŭ vilaĝo. La vertikaloj de la esplor-matrico reprezentas **atributojn**. Tiuj estas ĉioj, kioj rilatas al la esplorunuo: la mezuraĵoj kaj ĉiuj ecoj kiujn oni estas krome protokolinta.

Ekzemplo : Ĉe enketo pri la arbarmortado oni volas esplori en 3 malsamaj areoj. Super ĉiuj el ili oni metas kradon, kaj kiel esplorunuojn oni prenas la arbojn, kiuj situas

plej proksime al la kradopunktoj. La protokolitaj atributoj estas :

A1 : numero de la areo [kodita per 1, 2, 3]

A2 : arbo-specio [kodita per 1=piceo, 2=abio, 3=pino, 4=fago, 5=aliaj]

A3 : alto [indikita per metroj]

A4 : diametro en alto de 1.30 m [indikita per centimetroj]

A5 : sanstato [kodita per 0=sana, 1=malsaneta, 2=malsana, 3=malsanega, 4=morta]

Ĉe la opaj esplorunuoj la atributoj alprenas valorojn, kiuj kutime estas koditaj per nombroj. Tiuj atributo-valoroj (kaj ankaŭ la nombroj ilin kodantaj) nomiĝas **realigaĵoj**. Ĉe la supra ekzemplo povus okazi, ke ĉe esplorunuo 123 la atributo A1 havas la realigaĵon 2, A2 havas 4, A3 havas 7.43, A4 havas 9.5, kaj A5 havas 1. Tio signifas, ke la koncerna arbo estas malsaneta fago, 7.43 m alta kaj 9.5 cm dika, kiu kreskas en la 2a areo. La aro de valoroj, kiujn povas alpreni la realigaĵoj, estas la **skalo** de la atributo.

Inter la atributoj de la supra ekzemplo ekzistas gravaj malsimilecoj :
Unue : Ili malsamas laŭ la tipo de la skalo : Atributo nomiĝas **metrike skaligita**, se oni povas el du ĝiaj realigaĵoj formi senc-havan diferencon. En nia ekzemplo metrike skaligitaj estas la atributoj A3 kaj A4. Oni povas diri, ke la alto-diferenco inter arbo alta je 10 m kaj arbo alta je 12 m estas same granda kiom la alto-diferenco inter arbo alta je 7 m kaj arbo alta je 9 m, kaj ĝi estas duoble tiom granda kiom la diferenco inter arbo alta je 11 m kaj arbo alta je 12 m. Sed ne havas sencon diri, ke la sanstat-diferenco inter sana kaj malsana arbo estas tiom granda kiom inter malsana kaj morta, kaj duoble tiom granda kiom inter sana kaj malsaneta. Eĉ pli absurde estus diri, ke la speci-diferenco inter abio kaj fago estas same granda kiom inter piceo kaj pino kaj duoble tiom granda kiom inter pino kaj fago, aŭ ke la diferenco inter areo 1 kaj 3 estas la duobla kompare kun la diferenco inter areo 1 kaj 2. El tio sekvas, ke la atributoj A1, A2 kaj A5 ne estas metrike skaligitaj.

Sed por atributo A5 almenaŭ eblas kompari kaj vicordigi la realigaĵojn laŭ senc-hava maniero, ekz. laŭ la vico sana < malsaneta < malsana < malsanega < morta. Atributo, kies realigaĵoj povas esti tiamaniere vicordigata, nomiĝas **ordinale skaligita**. Tia vicordigo ne eblas por atributoj A1 kaj A2. Ĉe ili la komparo inter du realigaĵoj povas rezulti nur en la aserto “ili samas” aŭ “ili malsamas”. Atributo, kiu permesas almenaŭ tiun simplan komparon, nomiĝas **nominale skaligita**.

Oni povas vicordigi la skalo-tipojn laŭ la ordo : nominala < ordinala < metrika. En la fakterminaro ankoraŭ ne estas decidite, ĉu la metrika skalo estu konsiderata kiel speciala ordinala skalo aŭ ne-ordinala skalo (la samo validas por la rilato inter la ordinala kaj la nominala skaloj). Tial, se la skalo de atributo estas ja ordinala, sed ne plu metrika, oni por klareco nomu ĝin **nur** ordinale skaligita; se oni volas diri, ke la realigaĵojn oni povas senc-have ordigi, sen diri ion pri la diferenceblo, oni nomu la atributon **almenaŭ** ordinale skaligita. Grava ekzemplo por atributo skaligita nur ordinale estas la (alt-)lernejaj notoj: ĉar oni ne rajtas aserti, ke la diferenco inter “kontentige” kaj “bonege” samas la diferencon inter “kontentige” kaj “mankhave” (laŭ

la germana notosistemo). Se atributo povas alpreni nur 2 malsamajn realigaĵojn, ekz. 0 por viva kaj 1 por morta, tiam la diferencoj inter la skaloj estas nuligitaj.

La **dua** diferenco inter la atributoj konsistas el tio, ke la realigaĵoj de la atributoj 3 kaj 4 povas esti ajnaj nombroj almenaŭ en intervalo de la reela akso. Tiaj atributoj nomiĝas **kontinuaj**. Se la eblaj realigaĵoj de atributo situas dise, tiam la atributo nomiĝas **malkontinua**. Ekzemploj estas A1, A2 kaj A5, kiuj povas alpreni eĉ nur entjerojn. (Oni ne nomu ilin "diskretaj", tiu vorto signifas ion tute alian) Atributoj skaligitaj nur ordinale aŭ nur nominale estas kutime malkontinuaj. Ekzemplo por metrike skaligita malkontinua atributo estas la nombro de idoj (aŭ la nombro de ekzamenaj fiaskoj).

Per la **tria** diferenco distingiĝas atributo A1 disde la aliaj. Ĉe A1 la esploristo ne esploras la realigaĵon, sed scias ĝin antaŭe, ĝi estas ja **determinita** per la esplorplano. Kontraŭe, la realigaĵoj de atributoj 2,3,4,5 varias hazarde. Tiajn atributojn oni nomas **stokastaj**. En eksperimentoj, oni klopodas determini ĉiujn atributojn krom la celvariabloj, dum ke en enketoj la nombro de stokastaj atributoj povas esti malagrable granda.

Ofte oni lasas la variablojn ne tiel, kiel ili estas, sed oni kalkulas el ili novajn per **transformo**. Imagu, ke oni mezuris la koncentritecon de iu substanco en mol/l, kaj oni scias, ke por la statistika prilaborado pli taŭgas ties logaritmo, do oni aldonas al la daten-matrico novan vertikalon kiu enhavas la atributon : log (konc.) . Aŭ oni mezuris temperaturon en gradoj Celsius kaj poste volas transformi ĝin en gradojn Fahrenheit. Oni kalkulas la novan atributon el la malnova per : temp.[F] = 32 + 1.8 $*$ temp.[C]. Novajn atributojn oni povas kalkuli ne nur el po unu alia atributo, sed ankaŭ el pluraj. Se oni ekz. mezurintas la koncentritecon de K, Ca , kaj Mg en iu grundspecimeno, oni eble interesiĝas pri la kvociento [K] / ([Ca]+[Mg]) , kiu nun estas nova stokasta atributo [Proporcio inter alkali- kaj teralkali-metaloj].

Esplormatrico, kian mi estas priparolinta, devas nun esti prilaborata por ke oni akiru el ĝi la deziratajn informojn. La tutaĵon de la metodoj por tion fari oni povas disdividi en du partojn. Unue, oni devas prezenti la eble grandegan daten-amason en klara, komprenebla formo. La taŭgajn metodojn por akiri tion provizas la **deskripta* statistiko**. La metodojn por konkludi el nia datenaro al la pli granda aro, el kiu ĝi estas prenita, liveras la **konkluda statistiko**. Tiuj du branĉoj de la aplikata statistiko ne estas sendependaj unu de la alia, ĉar la prilaboro de la datenaro helpe de deskriptaj metodoj nepre antaŭas ĉiun konkludan proceduron.

III. Deskripta statistiko

La tasko de la deskripta statistiko estas prezenti la akiritan datenaron en bone komprenebla, **neni-okaze erariga** formo, gardante plejeble bone ĉiun gravan informon. Tiun prezentadon oni povas unue efektivigi per grafikaĵoj (kurboj, diagramoj, histogramoj kaj multaj aliaj); la metodaro por tia prezentado helpe de komputila programar-sistemoj multe evoluis en la lasta tempo. Tamen tio estas tro speciala kampo por enkonduka kurso [1], do ni ne okupiĝos pri tio. Alia, pli vaste uzata metodo

[1] Tion mi skribis en 1989, hodiaŭ la perkomputila produktado de grafikaĵoj apartenas al la plej bazaj teknikoj

por deskripti datenaron konsistas el tio, ke oni kalkulas el ĝi kelkajn nombrojn, kiuj plej bone karakterizu ĝin. Tiaj nombroj en la angla lingvo nomiĝas "statistics", oni do uzas por ili la saman vorton kiun por la scienco mem. Ĉar mi pensas, ke tia plursenceco de termino estas nepre evitenda, mi nomos tian nombron **adedo** (la vort-radiko venas el la araba lingvo). Ankaŭ la kalkulado de adedoj estas transformo, kiu nun ne plu limiĝas al la opaj horizontaloj, sed transformas la tutan datenmatricon samtempe.

III.a Lokaj adedoj

Komence mi volas priparoli adedojn, kiuj koncernas nur po unu atributon, do ne la interdependon inter du atributoj. Kutime, tiuj adedoj estas senc-havaj nur ĉe stokastaj variabloj. La plej gravaj el ili estas la **lokaj adedoj**. Tiuj indikas, kie sur la nombrorekto troviĝas la tipaj realigaĵoj de la atributo. La plej konata (kaj samtempe la plej senskrupule misuzata) el la lokaj adedoj estas la **aritmo**. Ĝi estas kalkulata simple per : $\bar{y} = \sum_{i=1}^{n} y_i/n$, kie $\bar{y}$ estas la aritmo, y_i - la realigaĵoj de la koncernata atributo kaj n - la nombro de esplorunuoj. La kaŭzo por la graveco de la aritmo estas, ke la sumo de la kvadratigitaj diferencoj inter la realigaĵoj kaj la aritmo estas minimumo :

$$\bar{y} = \min_{z \in \Re} \sum_{i=1}^{n} (y_i - z)^2$$

El tio sekvas, ke la aritmo estas senc-hava adedo nur, se la koncernata atributo estas metrike skaligita. Ĝi ne bone deskriptas vicon de ordinale skaligitaj realigaĵoj, ekz. lernejaj notoj, kvamkam ĝuste por tiuj ĝi tre ofte estas (mis-)uzata.

Ankaŭ ĉe metrike skaligitaj atributoj ĝi ne ĉiam liveras utilan informon, kaj ofte eĉ misinformas nekritikemajn homojn. Por doni ekzemplon por tia misinterpretado: En iu popularscienca libro pri prahistorio la aŭtoro skribas, ke en la ŝtona epoko la aritma vivodaŭro estis nur 18 jaroj - en tiu li pravas - kaj ke oni do devas konkludi, ke ne ekzistis mezaĝaj aŭ maljunaj homoj kiuj povintus transdoni sciojn kaj tradiciojn. Tio estas laŭ lia opinio la kaŭzo por la malrapidega kulturevoluo en la prahistorio.

Tio estas tute erara konkludo. La malaltan aritmon de la vivodaŭro kaŭzis ne la manko de maljunuloj, sed la altega infanmortemo. Se oni aritmas la vicon
(0 0 0 0 100), oni ricevas 20, do nombron, kiu situas malproksime de ĉiuj observitaj realigaĵoj. Eĉ pli ekstreman situacion ni trovas ĉe la aritmaj vivodaŭroj de arboj: Ekzistas arbospecioj, kiuj povas atingi aĝon de kelkmil jaroj. Sed la ega plejmulto de la ĝermintaj arbidoj jam dum la unua jaro formortas, ĉar sub la foliaro de la plenkreskaj arboj ili ne ricevas sufiĉe da lumo. Nur kiam mortas maljuna arbo, inter la arbidoj sub ĝi ekestas vetkreskado. Tiun vetkreskadon gajnas unu arbo, kiu nun efektive havas bonan ŝancon por longega vivo. Do, la aritma vivodaŭro de tia longvivema arbospecio estas - eble inter unu kaj du jaroj. Ankaŭ en tiu-ĉi okazo la aritmo ne liveras uzeblan informon. Por ke oni ekhavu aritmon kun senc-hava informo, estas unu kondiĉo ke la realigaĵoj proksimume situas simetrie ĉirkaŭ iu centro. Kaze de vivodaŭro tiu kondiĉo estas plenumita nur ĉe la vivodaŭro de homoj en landoj kun malgranda infanmortemo. Alikaze ni havas grandan amason de nombroj, kiuj preskaŭ egalas al 0 , malmultajn mezgrandajn kaj kelkajn grandajn, la distribuo de la realigaĵoj do estas ne simetria, sed oblikva.

Ekzistas ankoraŭa kaŭzo por malutiligi la aritmon. Ofte okazas, ke la plej granda parto de la realigaĵoj konsistas el nombroj, kiuj situas sufiĉe simetrie ĉirkau iu centro, sed kelkaj malmultaj situas malproksime de ĝi. Kiam oni kolektas la datenojn, povas okazi, ke oni hazarde ricevas nur la "normalajn" realigaĵojn , sed kelkfoje oni ricevas ankaŭ unu aŭ eĉ du el la "nenormalaj" . En tia kazo, la aritmo sufiĉe forte varias dependo de tio, ĉu ĝi hazarde enhavas "nenormalajn" realigaĵojn aŭ ne. Por doni ekzemplon : oni faras fiziologian eksperimenton kun aro da homoj, kaj povas okazi, ke unu el ili estas malsana kaj respondas nenormale. Aŭ oni enketas plantojn, el kiuj kelkaj malmultaj estas difektitaj de bestoj, sen ke oni rimarkis tion (eble okazis difekto de la radiko).

Tiaj "nenormalaj" realigaĵoj nomiĝas angle "outlier" = el-kuŝantoj, germane eĉ pli draste "Ausreißer" = forfuĝintoj. En Ilo mi nomas ilin **distantoj**, ĉar ili situas diste (tio estas malproksime) de la aliaj. La aritmo ne estas **robusta** kontraŭ distantoj, tio signifas, ke ĝi estas sufiĉe forte influata de ili, ĝis preskaŭa neuzeblo.

Resume: Por ke la aritmo estu efektive uzebla, devas esti plenumataj tri premisoj: unue : la respektiva atributo devas esti skaligita metrike ; due : la distribuo de la realigaĵoj devas esti simetria ; trie : inter la datenoj ne estu distantoj.

Ekzistas alia loka adedo, kiu kavas pli grandan aplikeblon, sed ĝis nun, malgraŭ sia taŭgeco, ne estas tre bone konata inter nefakuloj. Temas pri la **duilo**. Oni akiras ĝin jene: Oni vicordigas la realigaĵojn kaj prenas tiun en la mezo de la ordigita vico. Se la nombro de la datenoj estas para, oni prenas la aritmon de la du en la mezo. Du etaj ekzemploj :
(3 6 2 4 1 7 1) $\Rightarrow$ (1 1 2 $\underline{3}$ 4 6 7)$\Rightarrow$ 3
(5 3 1 1 3 6 8 5) $\Rightarrow$ (1 1 3 $\underline{3\ 5}$ 5 6 8) $\Rightarrow$ 4
La duilo donas bonan informon, se la atributo estas almenaŭ ordinale skaligita kaj la distribuo de la realigaĵoj ne estas tro oblikva; precipe ĝi estas tute ne afektata de distantoj. Kompreneble ĉe nominalaj skaligoj ĝi ne estas aplikebla. La koncepto de la duilo povas esti ĝeneraligata : Oni determinas en la ordigita vico de realigaĵoj tiun nombron, sub kiu situas kvarono de la valoroj, kaj tiun adedon oni nomas unua **kvarilo**. La nombro, sub kiu situas 3 kvaronoj estas do la tria kvarilo, kaj konsekvence la duilo estas samtempe la dua kvarilo. Aliaj similaj adedoj estas la 9 dekiloj kaj la 99 centiloj. Por tiuj adedoj ekzistas komuna termino, kiu estas **plurilo**. La unua dekilo estas samtempe la 10%-a plurilo, same la tria centilo la 3%-plurilo. Aparte gravaj en la statistiko estas la 5%- kaj 2.5%- resp. la 95%- kaj 97.5%-pluriloj. Helpe de pluriloj ni povas karakterizi ankaŭ oblikvajn distribuojn.

III.b Dispersaj adedoj

Por bone karakterizi aron da realigaĵoj, ne sufiĉas loka adedo, ĉar ankaŭ simetrie aranĝitaj datenoj povas malsami ne nur konsidere la lokon de la aritmo aŭ de la duilo, sed ankaŭ per tiu, kiom ili dispersiĝas ĉirkaŭ la loka adedo. Eta ekzemplo : La (neordigitaj) vicoj (33 42 37 42 31) kaj (22 3 72 37 51) havas la saman aritmon kaj la saman duilon (ambaŭ estas 37), tamen ili distingiĝas per tio, ke en la unua vico la nombroj situas multe pli proksime unu al la alia. Tiun interan proksimecon oni deskriptas per **dispersaj adedoj**. La plej konata el ili estas la **varianco**. Ĝi estas difinita per jena

formulo

$$var(y) = SQy/(n-1) = [\ \sum_{i=1}^{n}(y_i - \bar{y})^2\ ]\ /\ (n-1),$$

kie SQ signifas "sumo de kvadratoj (de la devioj)". La varianco estas do (preskaŭ) la aritmo de la kvadratigitaj distancoj inter la realigaĵoj kaj la aritmo. Por plifaciligi la kalkuladon, oni transformas tiun ekvacion al

$$SQy = \sum_{i=1}^{n} y_i^2 - (\ \sum_{i=1}^{n} y_i\ )^2\ /\ n$$

Por apliki tiun formulon : La varianco de la unua el la supre prezentitaj vicoj estas :
$var(y) = [\ 33^2 + 42^2 + 37^2 + 42^2 + 31^2\ -\ (33+42+37+42+31)^2\ /\ 5]\ /\ 4 =$
$[1089 + 1764 + 1369 + 1764 + 961 - 185^2\ /\ 5]\ /\ 4 = [6947 - 6845]/4 = 102/4 = 25.5$
La varianco de la dua vico (bv. mem kalkuli) estas 700.5

La varianco ne taŭgas en kazoj, en kiuj la aritmo mem estas ne informiva, distantojn ĝi toleras eĉ malpli ol la aritmo. Ofte oni uzas anstataŭ la varianco ties duan radikon. Tiu adedo nomiĝas laŭ PIV "varianca devio". Laŭ mia opinio tiu termino tute ne taŭgas, ĉar ne temas pri la devio de la varianco. Mi preferus la terminon **ordinara devio**, kiu similas la kutimajn terminojn en la etnaj lingvoj (standard deviation, Standardabweichung). Alia dispersa adedo, kiu taŭgas ankaŭ en kazoj, en kiuj la varianco ne estas informiva, estas la distanco inter difinitaj pluriloj, plej ofte uzatas la distanco inter la unua kaj la tria kvarilo, la **interkvarila distanco**.

Krom lokaj kaj dispersaj adedoj ekzistas ankaŭ formaj adedoj, ekzemple adedoj de nesimetrio. Sed pri tiuj ni ne okupiĝas dum nia kurso.

El la ĝis nun priparolitaj adedoj oni povas kalkuli pluajn. Plej grava el tiuj estas la kvociento el ordinara devio kaj aritmo. Tiu kvociento nomiĝas **variada koeficiento**. Por la supraj vicoj la variadaj koeficientoj estas 0.136 kaj 0.715.

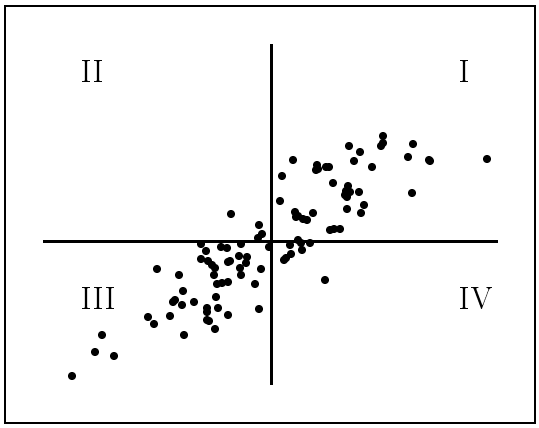
III.b Interdependecaj adedoj

Nun ni volas okupiĝi pri kelkaj - vere nur kelkaj - ekzempleroj el la grandega amaso de tiuj adedoj, kiuj deskriptas la rilaton inter pluraj atributoj. Kiu adedo estas pokaze vere informiva, tio dependas inter alie de tio, kiel la koncernaj atributoj estas skaligitaj, ĉu ili estas kontinuaj aŭ ne, kaj ĉu temas pri du stokastaj variabloj aŭ pri unu stokasta kaj unu determinita.

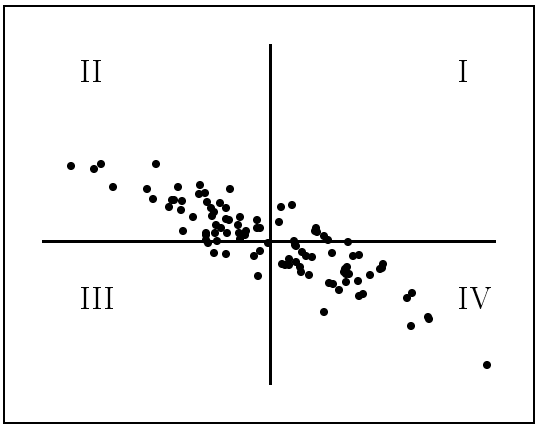
Unue ni volas pritrakti la kazon de du metrike skaligitaj stokastaj atributoj. Ili aŭ estu kontinuaj aŭ almenaŭ disponu pri multaj malsamaj realigaĵoj. Ekzemplo por tia atributo-paro estas la koncentritecoj de iuj substancoj A kaj B en la grundo aŭ en la sango de iu vivestaĵo. En la figuroj 1 ĝis 4 estas grafike prezentataj aroj da punktoj, kies koordinatoj reprezentas la du atributojn. Oni tuj vidas, ke la 4 figuroj montras tute malsamajn interrilatojn inter la koncernataj atributoj. En figuro 1 ni trovas, ke (mal)grandaj realigaĵoj de atributo 1 estas kutime (kvankam ne ĉiam!) ligitaj al (mal)grandaj realigaĵoj de atributo 2. Inverse estas en figuro 2 : Ĉi-tie la grandaj realigaĵoj de unu atributo koincidemas kun malgrandaj de la alia. En figuro 3 tute mankas iu rilato inter la du atributoj. Plej komplike estas la situacio montrata per figuro 4. Ĉi-tie la plej grandaj **kaj** la plej malgrandaj realigaĵoj de atributo 1 ligiĝas al

grandaj realigaĵoj de atributo 2, dum ke la mezgrandaj realigaĵoj de atributo 1 ligiĝas al la malgrandaj realigaĵoj de atributo 2.

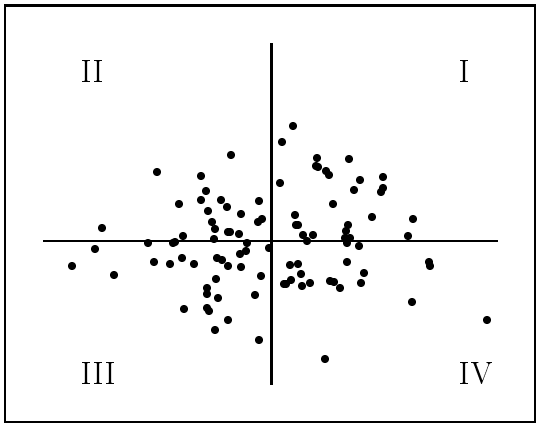
FIGURO 1



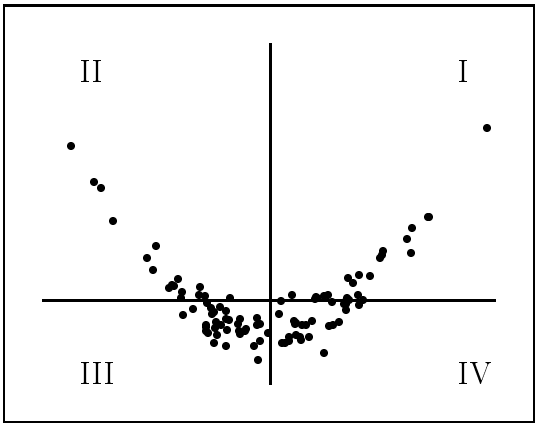
FIGURO 2



FIGURO 3



FIGURO 4



Por mezuri la interrilaton inter du tiaj atributoj, oni kutime uzas la **kovariancon**, kies formulo estas

$$kov(x,y) = SPxy \ / \ (n-1) = [\ \sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})\ ] \ / \ (n-1),$$

kie SP signifas "sumo de produtoj (de la devioj)". Por kompreni la sencon de ĉi-tiu formulo oni rigardu denove la figurojn 1 ĝis 4. En ili la aritmo de la unua atributo, x, estas indikita per vertikala streko, la aritmo de y - per horizontala streko. Tiel-ĉi ni ricevas kvar kvadrantojn. La realigaĵoj situantaj en kvadranto I estas super-aritmaj kaj rilate atributon 1 kaj rilate atributon 2. Por tiuj punktoj la faktoroj $(x_i - \bar{x})$ kaj $(y_i - \bar{y})$, kaj sekve ankaŭ iliaj produtoj estas pozitivaj. Por la punktoj en kvadranto III ambaŭ faktoroj estas negativaj, sekve ankaŭ ili liveras produtojn pozitivajn. Kontraŭe, por la

punktoj en la kvadrantoj II kaj IV unu el la faktoroj estas pozitiva, la alia negativa, la produto estas do negativa. El tio sekvas, ke pozitivan kovariancon havas punktaro, kies elementoj troviĝas plejgrandparte en kvadrantoj I kaj III, do en kiu (mal)super-aritmaj realigaĵoj de atributo 1 koincidemas kun (mal)super-aritmaj realigaĵoj de atributo 2 (vidu figuron 1). Negativan kovariancon havas punktaro, kie la rilatoj inter la atributoj estas inversaj (vidu figuron 2). Se la grandoj de la du realigaĵoj varias tute sendepende, tiam la punktoj troviĝas egale ofte en ĉiuj kvar kvadrantoj, do la kovarianco alproksimiĝas al 0 (vidu figuron 3). Aliflanke, se la kovarianco estas (preskaŭ) nulo, tiam oni ne rajtas konkludi el tio, ke la du koncernaj atributoj nepre havas neniun interrilaton. Kontraŭan ekzemplon montras figuro 4, kie la kovarianco estas preskaŭ nulo, tamen la atributoj sufiĉe forte interrilatas.

Por plifaciligi la kalkuladon, oni transformas ilian difinan ekvacion al la formo :

$$kov(x,y) = SPxy/(n-1) = [\ \sum_{i=1}^{n} x_1 * y_i\ -\ (\sum_{i=1}^{n} x_i) * (\sum_{i=1}^{n} y_i)/n\ ]\ /\ (n-1)$$

Por doni etan ekzemplon: Por la paro-vico ((3 , 5) , (4 , 7) , (1 , 4) , (5 , 7) , (3 , 6)) ni ricevas

$$kov(x,y) = [3*5+4*7+1*4+5*7+3*6-(3+4+1+5+3)*(5+7+4+7+6)/5]\ /\ 4 =$$

$$[100 - 16 * 29/5]/4 = 1.8$$

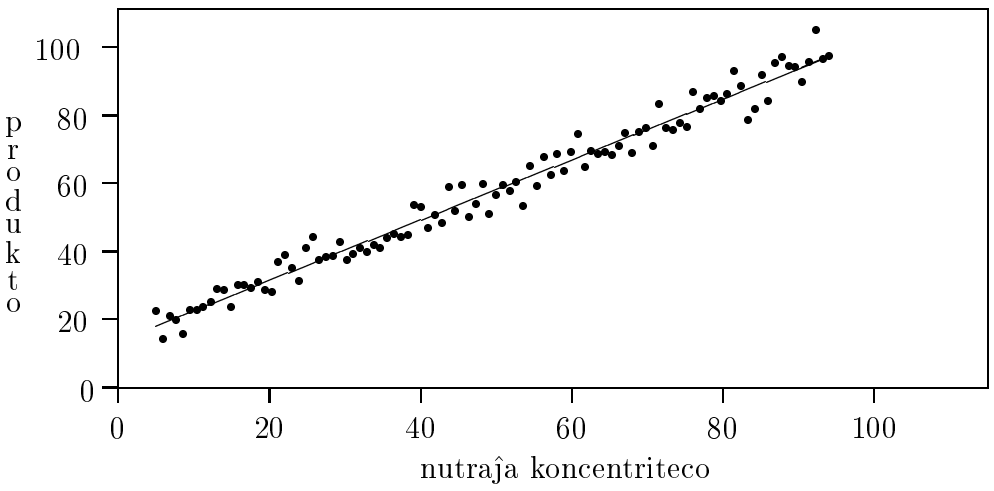
Kutime oni normas la kovariancon, dividante ĝin per la produto de la du ordinaraj devioj. Tiu nova adedo nomiĝas (ordinara aŭ Pearson'a) **korelaci-koeficiento** $r_{x,y}$. Ĝi havas ĉiam la saman signumon kiel la koncerna kovarianco, ĉar la denominatoro ja estas ĉiam pozitiva, sed ĝi estas normita tiel, ke ĝi alprenas valorojn nur inter -1 kaj +1. Ju pli ofte grandaj valoroj de unu atributo koincidas kun (mal)grandaj valoroj de la alia, des pli la korelaci-koeficiento alproksimiĝas al +1 resp. -1, kaj se **tia** interrilato ne ekzistas, ĝi alproksimiĝas al 0.

La korelaci-koeficiento de la supra paro-vico estas 1.8 / (1.4832∗1.3038) = 0.9308. La punktaroj en la figuroj 1 ĝis 4 liveras la respektivajn korelaci-koeficientojn :
fig.1: 0.87 ; fig.2: -0.88 ; fig.3: 0.08 ; fig.4: 0.05.

Ni transiru nun al alia sed simila adedo. Ni havu unu atributon x el metrika skaligo, kiu povas esti aŭ stokasta aŭ determinita, kaj alian atributon y , ankaŭ el metrika skaligo, kiu nepre estu stokasta. Ekzemple: Ni priesploras la interrilaton inter grundaj nutraĵelementoj kaj planta produkto. Oni povas imagi du malsamajn esplorplanojn: Aŭ oni aldonas diversajn kvantumojn al la opaj esplor-parceloj (ĉi-okaze variablo x estas determinita), aŭ oni ne traktas la parcelojn diversmaniere, sed mezuras la nature disponeblajn nutraĵelementojn (ĉi-okaze x estas stokasta). En ambaŭ kazoj ni povas per grafikaĵo demonstri, kiel y (la produkto) dependas de x (la nutraĵo). Ofte tiu interrilato estas proksimume lineara, tio signifas, ke oni povas desegni rekton trairantan la punktaron, kiu proksimume adaptiĝas al la punktaro (Vidu figuron 5). Tiu adaptiĝanta rekto nomiĝas **regresi-rekto**. Atributo x estas la **sendependa** aŭ **determinanta**, y la **dependa** aŭ **determinita**. Kiel ĉiu rekto, la regresi-rekto estas plene difinita per

du koeficientoj, la intercepto b_0 (tranĉejo kun la ordinato) kaj la dekliveco b_1 (tangento de la dekliva angulo), kiuj difinas y kiel $y = b_0 + b_1 * x$. b_0 kaj b_1 estas la regresi-koeficientoj, kiuj bone taŭgas kiel adedoj por deskripti, kiel y dependas de x.

FIGURO 5



Por trovi la plej bone adaptiĝantan rekton, oni difinas la adapto-bonecon de donita rekto jene: Estu r_i la **reziduo** de la i. y-valoro, tio estas la vertikala distanco inter y_i kaj la rekto: $r_i = y_i - (b_0 + b_1 * x_i)$. Tiun rekton, por kiu la sumo de la kvadratigitaj reziduoj estas minumumo, oni elektas kiel plej bone adaptiĝantan.

$$(b_0, b_1) = \min_{z_0, z_1 \in R} \sum_{i=1}^{n} r_i^2 = \min_{z_0, z_1 \in R} \sum_{i=1}^{n} (y_i - (z_0 + z_1 * x_i))^2$$

Per simpla minimaks-kalkulado oni povas nun determini la koeficientojn de ĉi-tiu rekto, kaj oni ricevas la formulojn:

$$b_1 = SPxy/SQx \ ; \ b_0 = \bar{y} - b_1 * \bar{x}$$

(signifon kaj kalkuleblon de SP kaj SQ vidu supre). La signumo de b_1 egalas la signumon de la koncerna korelaci-koeficiento. Ankaŭ la disperso de la punktoj ĉirkaŭ la regresia rekto estas informiva adedo. Oni deskriptas ĝin per la ĵus enkondukita sumo de la kvadratigitaj reziduoj, dividita per la nombro de la punktoj minus la nombro de la regresi-koeficientoj, do ĉi-kaze minus 2 . Tiu adedo nomiĝas **varianco ĉirkaŭ regresio** aŭ **rest-varianco**. Plej simple oni kalkulas ĝin per la formulo :

$$s^2_{resto} = [SQy - b_1 * SPxy] \ / \ (n - 2).$$

Plej ofte iu atributo y dependas ne nur de unu alia atributo, sed de pluraj, diru de $x_1, ..., x_m$. La punktoj $(y_i\ , x_{1,i}, .., x_{m,i})$ situas en $(m+1)$-dimensia spaco, kaj oni povas kalkuli m-dimensian hiperebenon, kiu adaptiĝas plej bone al la y-valoroj.

$$y_i = b_0 + b_1 * x_{1,i} + ... + b_m * x_{m,i}$$

Tian regresion oni nomas **multobla regresio**. Oni do regresias la dependan (determinitan) y sur la sendependajn (determinantajn) $x_1, .., x_m$. Por kalkuli la restvariancon, oni nun dividas la sumon de la kvadratigitaj reziduoj per la nombro de punktoj minus $(m+1)$.

Ekzemplo : En la forstmastrumado oni devas scii, kiom da kubmetroj da ligno enhavas iu arbaro, por ne forhaki tro. Kompreneble la ekzakta determinado de la volumeno estas tre temporaba kaj eblas nur ĉe hakita arbo. Sed oni scias, ke la volumeno bone rilatas al kelkaj pli facile mezureblaj atributoj, ekzemple al la grando kaj al la diko. Malbonŝance ne ekzistas kontentige bona regresio inter la volumeno kaj nur unu alia atributo, sed oni ja povas determini la volumenon helpe de pluraj atributoj samtempe, ekzemple per $\ln(v) = b_0 + b_1 * \ln(d_{1.3}) + b_2 * \ln(a) + b_3 * \ln(d_i)$, kie v estas la volumeno, $d_{1.3}$ - la diametro en la norma alteco de 1.3 m, d_i - la diametro en alia alteco, (ekz. en 30 m), a - la alteco de la arbo. Ĉi-tie ĉiuj atributoj enirantaj la regresi-ekvacion estas transformaĵoj de observitaj atributoj: $y = \ln(v)$, $x_1 = \ln(d_{1.3})$, $x_2 = \ln(a)$, $x_3 = \ln(d_i)$. Por unufoje determini la koeficientojn, oni fakte devas haki kelkajn arbojn kaj precize mezuri ilin. Helpe de tiuj mezuraĵoj oni kalkulas la koeficientojn, kaj se la restvarianco estas sufiĉe malgranda, oni povas poste determini la arban volumenon per la mezurado de la tri atributoj "alto de la arbo", "diametro en alto de 1.3m" kaj "diametro en alto de 30 m".

Ofte la rilato inter la determinantaj variabloj $x_1, .., x_m$ kaj la determinata variablo y ne estas lineara (komparu figuro 4). Ĉi-kaze oni povus deskripti y per polinomo, ekzemple per $y = b_0 + b_1 * x + b_2 * x^2 + b_3 * x^3$. Tia polinoma regresio estas nenio alia ol speciala multobla regresio, ĉar ni regresias y sur la variablojn x, x^2 kaj x^3, el kiuj x estas mezurita, x^2 kaj x^3 aldonitaj kiel transformaĵoj. Ekzemplo el la forstmastrumado estas: $v = b_0 + b_1 * d^2 * a + b_2 * d * d_i * a + b_3 * a^2$. Ĉi-tie ni havas nekompletan polinomon de tria grado kun tri variabloj.

Dum la analizado de interrilatoj ekestas fojfoje jena problemo: Ni havas la stokastajn atributojn y_1 kaj y_2, kiuj ambaŭ korelacias kun tria atributo x. El tio sekvas, ke simple pro tiu fakto ekzistas kutime ankaŭ korelacio inter y_1 kaj y_2. Sed oni volas scii, ĉu ekzistas ankaŭ korelacio inter y_1 kaj y_2, kiu **ne** sekvas el la influo de x. Ekzemple estu konate, ke ĉe iu planto la kreskorapido y_1 kaj la aktiveco y_2 de iu hormono ambaŭ dependas de la temperaturo x. Ekestas nun la demando, ĉu ekzistas korelacio inter kreskorapido kaj hormonaktiveco ankaŭ sendependa de la temperaturo. Por respondi al tiu demando, oni regresias y_1 kaj y_2 sur x kaj kalkulas por ĉiu esplorunuo la du reziduojn. Tiam oni kalkulas la korelaci-koeficienton inter tiuj du reziduoj. Tiu nomiĝas **parta korelaci-koeficiento** inter y_1 kaj y_2 post forigo de la influo de x, mallonge $r_{y_1,y_2.x}$. Same oni povas kalkuli la korelacion inter y_1 kaj y_2 post forigo de la influo de pluraj determinantaj variabloj kaj ekhavas $r_{y_1,y_2.x_1,..,x_m}$.

IV. Konkluda statistiko

La konkluda statistiko rigardas la donitajn esplorunuojn kiel sub-aron el pli granda, eble nefinia aro, kaj ĝi klopodas konkludi el la ecoj de tiu-ĉi sub-aro al la ecoj de la aro, el kiu la esplorunuoj estas prenitaj. La superordigita aro, kies ecojn oni intencas esplori, nomiĝas **populacio**, la aro de esplorunuoj - **specimeno**. Por ke oni fakte povu akiri validajn informojn pri la populacio, oni devas plenumi kelkajn regulojn dum la elektado de la specimeno. Por la plej simplaj statistikaj analizoj (nur tiajn ni pritraktos ĉi-tie) oni prenu **aleatoran** specimenon. La difino de aleatora specimeno estas laŭ PIV: "Specimeno elektita tiamaniere , ke ĉiu kombinaĵo de donita nombro de unuoj havas la saman probablon esti elektita (= ∼a specimenado)". La aro de la realigaĵoj de la stokastaj atributoj nomiĝas **samplo**, se la specimeno estas prenita laŭ statistike pravigebla metodo, ekz. aleatore. En la etnaj lingvoj ne ekzistas tiu diferencigo inter specimeno kaj samplo, kiu estas difinita en PIV, ankaŭ ne en la angla, el kiu la koncernaj vortradikoj estas prenitaj. Ankoraŭ ne estas certe, ĉu tiuj konceptoj fakte enradikiĝos en la internacilingva fakterminaro.

Por doni kelkajn ekzemplojn por aleatora specimenado el finia kaj nefinia populacioj: Supozu, ke oni interesiĝas pri iu atributo de la arboj de difinita arbaro (alto, sanstato, lastjara kresko). Ĉar oni ne povas priesplori ĉiujn arbojn de ĉi-tiu populacio, oni elektas specimenon. Kiel specimeneroj ne estu prenataj ekz. nur arboj kreskantaj laŭlonge de la vojoj, aŭ arboj kreskantaj en la sama sub-areo de la arbaro, sed oni prenu ilin hazarde el la tuta arbaro. Se oni faras demoskopian opini-esploron, oni ne prenu kiel specimenon la homojn, kiujn oni renkontas sur iu placo aŭ strato, sed oni elektu ilin lotece el adresaro, kiu enhavas la **tutan** populacion. Se oni prenas la adresojn el telefonlibro, tiam la koncerna populacio, por kiu la konkluda statistiko liveras informojn, estas la aro de la telefonposedantoj, **ne** la aro de ĉiuj (adoltaj) homoj de la priesplorita regiono!

La celo de scienca, specife naturscienca esploro plej ofte ne estas ekkonoj pri iu difinita finia populacio, sed la malkovro de scienca leĝo. Se ni ekzemple ekzamenas, ĉu iu substanco X rapidigas la kreskon de blankaj abioj, tiam interesas nin la ebla efikivo de tiu substanco sur ĉiajn ekzemplerojn de la specio "blanka abio", ne nur la nun vivantajn, sed ankaŭ la iam vivintajn aŭ vivontajn aŭ imageble vivuntajn. La esplorata populacio estas do nefinia, oni nomas ĝin ankaŭ "hipoteza". Por ke la specimeno estu aleatora, oni ne rajtas preni nur plantojn inter si parencajn aŭ selektitajn laŭ iu atributo (alto, konkurivo ktp.).

IV.a Bazaj konceptoj de la probablo-teorio

La konkludoj de la specimeno al la populacio neniam estas certaj, sed ili validas je difinita **probablo**. En la statistiko, la koncepto de la probablo havas pli precizan sencon ol en la ĉiu-taga lingvaĵo. En la matematika probablo-teorio la probablo enkondukatas per aksiom-sistemo. Ĉar en enkonduka kurso ni ne povas profundiĝi en la matematikon, ni difinu ĝin ĉi-tie tiamaniere, kiel estas plej taŭge por la aplikanto de la statistiko. Tiu difino tekstas jene: Estu la okazoj OK_i ($i = 1, .., n$ aŭ $i = 1, .., \infty$) eblaj rezultoj de esploro. Tiam la probablo por OK_i, simbole $P(OK_i)$, estas la proporcia ofteco de la apero de tiu okazo, se ni observas grandegan serion de ripetoj: $P(OK_i) = \lim_{n\to\infty} ofteco(OK_i)/n$. Simplega ekzemplo: La esploro estu kubĵetado,

kaj OK_i estu la okazo "akiro de pli ol 4 poentoj". Tiam $P(OK_i) = 1/3$, kondiĉe ke la ĵetkubo estas ideala. Se oni parolas pri okazoj sen ke oni povas imagi eĉ teorie serion da ripetoj, oni evitu la vorton "probablo" kaj prefere uzu "supozeblo", ekz.: "Supozeble tiu fiulo malfeliĉigas sian amikinon". (Ankaŭ la vorto "verŝajno" havas specialan signifon en la statistiko, kiu tamen ne estos traktata ĉi-tie.)

Du okazoj OK_1 kaj OK_2 nomatas **sendependaj** unu de la alia, se la probablo, ke ambaŭ realiĝas, egalas al la produtoj de iliaj opaj probabloj, do se $P(OK_1$ kaj $OK_2) = P(OK_1) * P(OK_2)$. Se oni ekzemple ĵetas ludkubon dufoje, tiam la okazoj "6 poentoj je unua ĵeto" kaj "6 poentoj je dua ĵeto" ne estas interdependaj; ambaŭ havas probablon egale al 1/6, kaj ilia kajaĵo (t.e. $OK_1 \wedge OK_2$) havas probablon de 1/36. Kontraŭe, se OK_1 resp. OK_2 estas la okazoj, ke vi resp. via edz(in)o malsanos je 1990-01-01, tiam $OK_1 \wedge OK_2 > OK_1 * OK_2$. OK_1 kaj OK_2 ne estas sendependaj, ĉar familianoj ofte infektas unu la alian.

Ofte oni povas karakterizi okazon per nombro, ekz. per realigaĵo de stokasta atributo aŭ adedo de datenaro. Tiam la probabloj, je kiuj realiĝas la opaj nombroj, sekvas stokastan **distribuon**. Se stokasta variablo y sekvas distribuon D, oni simboligas tion per: $y \sim D$. Tian distribuon oni povas deskripti per funkcionoj*, el kiu la plej grava estas la **distribuo-funkciono**: Estu D la distribuo de la realigaĵoj de iu aro da okazoj (ekz. estu okazo i : akiro de i poentoj ĉe kubĵeto). Tiam ties distribuo-funkciono $DIF[D]$ estas difinita per : $DIF[D]\,(i) = P(r(D) \leq i)$, do la probablo por tio, ke realigaĵo el D estas malplieda* ol la funkciona argumento. Ĉe la ekzemplo de la kubĵetado ni ricevas:
$DIF[D]\ (i < 1) = 0$; $DIF[D]\ (1 \leq i \leq 6) = i/6$; $DIF[D]\ (i > 6) = 1$,
kie $i \in \aleph$ (aro de pozitivaj entjeroj);

Alia funkciono por karakterizi distribuon estas la **probablo-funkciono** $PRF[D]$, kiu indikas, por ĝia argumento i, la probablon, ke la realigaĵo egalas al i : $PRF[D]\,(i) = P(\,r(D) = i\,)$. Ĉe la kubĵetado ni havas: por $i = 1, .., 6$: $PRF[D](i) = 1/6$; alikaze $PRF[D](i) = 0$

Por kontinua distribuo (tio estas distribuo de kontinua atributo) la probablo-funkciono konstante egalas al 0, ĉar por ĝi validas:

$$\forall x : P(r(D) = x) = \lim_{\varepsilon \to 0} P(x - \varepsilon < r(D) < x + \varepsilon) = 0.$$

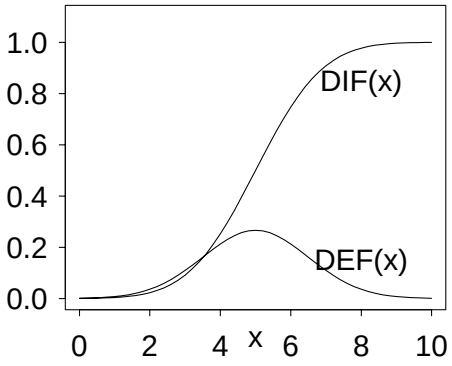
Tial la probablo-funkciono ne taŭgas por karakterizi kontinuajn distribuojn, anstataŭe oni prenas por ili la **denso-funkcionon** DEF[D], kiu estas la derivaĵo de la distribuo-funkciono (vidu malsupre la ekzemplon de la gaŭsa distribuo).

Ankaŭ populacioj kaj distribuoj havas adedojn. Ni ne rigardu ĉi-tie ties precizajn matematikajn difinojn, sed ni difinu ilin jene: Se la populacio estas finia, la adedo havas la saman signifon kiel ĉe la jam priparolitaj datenaroj; ali-kaze la populacia adedo estas la koncerna adedo el grandega aleatora samplo prenita el la populacio. Kutime la populaciaj kaj samplaj adedoj havas la saman nomon; la ununura escepto estas la aritmo, kies populacia ekvivalento nomiĝas **ekspekto**. Aparte gravaj populaciaj adedoj estas la pluriloj. La plurilo-funkciono $PLF[D]$ estas ĝuste la inverso de la distribuo-funkciono: $PLF[D]\ (\gamma) = x \iff DIF[D](x) = \gamma$.

Ni nun pritraktu du tipojn de distribuoj, kiuj aparte gravas en la aplikata statistiko. Estu $OK_1, .., OK_n$ serio de n ne interdependaj okazoj, kiuj liveras la rezulton "1" kun probablo p kaj la rezulton "0" kun probablo $1-p$. Ekz. OK_i povus esti ĵetado de monero, kie la apero de la nombro kodatas per "1", aŭ OK_i povus esti la rezulto de ĝermiga eksperimento, farita per n semoj, kie "1" signifas, ke semo i sukcese ĝermis. La probablo por tio, ke ekzakte k el la n OK_i liveris "1" (kaj sekve $n-k$ liveris "0"), do la probablo-funkciono por k, donitaj n kaj p, estas $PRF(k|n,p) = \binom{n}{k} * p_k * (1-p)^{n-k}$ por entjero k inter 0 kaj n, kie $\binom{n}{k}$ signifas $n!/(k! * (n-k)!) = \prod_{j=1}^{k}(n-j+1)/j$.

Distribuo kun tiu probablo-funkciono nomiĝas **dunomiala distribuo**, simbole $BIN(n,p)$, kie n estas pozitiva entjero kaj p reelo inter 0 kaj 1. La grandoj n kaj p estas la **parametroj** de ĉi-tiu distribuo. Ĉiam validas, ke distribuo estas unusence determinita per la indiko de sia tipo (ekz. dunomiala) kaj la valoroj de la parametroj. La parametroj de distribuo determinas ankaŭ ties adedojn, ekz:
ekspekto $(BIN(n,p)) = n * p$, varianco $(BIN(n,p)) = n * p * (1-p)$.

FIGURO 6



DIF kaj DEF de N(5,2.25)

La plej konata inter la kontinuaj distribuoj estas la **gaŭsa distribuo** $N(\mu, \sigma^2)$ kun la denso-funkciono

$$DEF[N(\mu,\sigma^2)](x) =$$

$$1/\sqrt{2 * \pi * \sigma^2} * \exp((x-\mu)^2/(2*\sigma^2))$$

Ĝiaj parametroj egalas la ekspekton kaj la variancon: ekspekto$(N(\mu,\sigma^2)) = \mu$ kaj varianco$(N(\mu,\sigma^2)) = \sigma^2$

La gaŭsa distribuo estas simetria kun unu kulmino; figuro 6 montras la denso- kaj distribuo-funkcionojn por N(5,2.25). Oni povas esprimi ajnan gaŭsan distribuon per ties normaĵo N(0,1): Se $y \sim N(\mu,\sigma^2)$, tiam $(y-\mu)/\sigma \sim N(0,1)$. La funkcionajn valorojn de la denso- kaj distribuo-funkcionoj de N(0,1), kaj sekve ankaŭ ties plurilojn, oni povas facile kalkuli per komputiloj aŭ ĉerpi el statistikaj tabeloj. La plej gravaj ĝiaj pluriloj estas: $PLF[N(0,1)](0.95) = 1.645$ kaj $PLF[N(0,1)](0.975) = 1.96$. Ĉar N(0,1) estas simetria ĉirkaŭ 0, ni ekhavas $DIF[N(0,1)](-x) = 1 - DIF[N(0,1)](x)$ kaj sekve $PLF[N(0,1)](\gamma) = -PLF[N(0,1)](1-\gamma)$, ekz. $PLF[N(0,1)](0.025) = -PLF[N(0,1)](0.975) = -1.96$.

Multaj kontinuaj atributoj estas almenaŭ proksimume gaŭse distribuitaj, tial tiu distribuo ludas gravan rolon en la aplikata statistiko. Eĉ malkontinuaj distribuoj kelkfoje alproksimiĝas al la gaŭsa, ekz. se $n*p*(1-p) > 9$,
tiam $BIN(n,p) \approx N(n*p, n*p*(1-p))$. Do, se $k \sim BIN(n,p)$ kun sufiĉe granda n, tiam $(k - n*p)/\sqrt{n*p*(1-p)} \sim N(0,1)$

IV.b Punkt-stimado

La plej simpla el la taskoj de la konkluda statistiko estas la **stimado** (pli precize: la punkt-stimado) de populaciaj adedoj kaj parametroj: Helpe de **stimo-funkciono** oni kalkulas el la samplo nombrojn, kiuj plejeble bone indikas la proksimuman valoron de la koncerna populacia grando. Oni simboligas stimaĵon per tegmento metita super la simbolon de la stimita grando. Ekz. por $BIN(n,p)$ ni ricevas: $\hat{p} = k/n$. La adedoj el samploj estas kutime konsideritaj kiel stimaĵoj por la koncernaj populaciaj adedoj, ekz. la aritmo kiel stimaĵo por la ekspekto. Ĉar la stimaĵojn oni kalkulas el hazarde fluktuantaj samploj, ili estas stokastaj variabloj, kiuj havas siajn distribuojn kun ekspekto kaj varianco. Por ke stimo-funkciono estu taŭga, ĝi devas, se eble, plenumi kelkajn postulojn. El ili estas plej gravaj la **ekspekto-fidelo**, la **varianc-minimumigo** kaj la **konsistento**. Stimado estas ekspekto-fidela, se la ekspekto de la stimaĵo estas la stimata populacia grando mem. Por la dunomiala distribuo oni povas fakte montri, ke ekspekto$(\hat{p} = k/n) = p$. Ankaŭ la aritmo kaj la samplo-varianco estas ekspekto-fidelaj stimaĵoj de la korianc-minimumigo kaj koncernaj populaciaj adedoj. Se oni difinintus ${}^*var(y) = SQy/n$ (anstataŭ $SQy/(n-1)$), tiam la sampla varianco **ne** estus ekspekto-fidela stimaĵo de la populacia varianco. La je unua aspekto stranga denominatoro n-1 estas motivita per tio, ke oni volis ekhavi adedon, kiu estu samtempe ekspekto-fidela stimaĵo. Stimo-funkciono nomiĝas konsistenta, se kun kreskanta samplo-amplekso ĝi nefinie alproksimiĝas al la stimata valoro. Finfine, se oni havas plurajn ekspekto-fidelajn kaj konsistentajn stimo-funkcionojn, oni prenas tiun kun la plej malgranda varianco.

IV.c Statistika testado

IV.c.1 Koncepto de statistika testo

Tre ofte okazas, ke la esploristo havas supozon pri populacia parametro, kaj li volas ekzameni, ĉu lia supozo estas pravigebla. Tion li povas esplori per **statistika testo**. La procedmanieron de tia statistika testo mi volas unue deskripti per ekzemplo, en kiu rolas la ĵus priparolita dunomiala distribuo. Estu konate, ke la ĝermo-probablo de glanoj (ĝermigita sub difinitaj cirkonstancoj) estas 45%. Oni nun volas esplori, ĉu la kemia substanco X efikas sur la ĝermivo. Por ekzameni tion, oni traktas, ni diru 25 glanojn per la substanco X kaj observas, kiom de la traktitaj glanoj ĝermas. La distribuo de la stokasta variablo $k =$ "nombro de la ĝermintaj glanoj" estas distribuita dunomiale kun $n = 25$ kaj p nekonata. Pri tiu nekonata, sed koninda parametro oni formulas paron da hipotezoj, nome la **testhipotezon** (kelkfoje ankaŭ nomata nulhipotezo) H_0, kiun oni provos malpruvi, kaj la **alternativ-hipotezon** H_1, kiun oni devas akcepti, se oni fakte sukcesos refuti H_0'on. En nia ekzemplo la hipotezo-paro konsistas el: $[H_0 : p = 0.45]$ kaj $[H_1 : p \neq 0.45]$. La ideo de statistika testo estas, ke oni malakceptu H_0'on (kaj sekve akceptu H_1'on), se la esplor-specimeno liveras rezulton, kiu havas tre malgrandan probablon sub H_0 (tio signifas: se la testhipotezo estas vera),

sed pli grandan probablon sub H_1. Por ke ni povu facile indiki tiujn probablojn, ni bezonas **test-adedon**, kies distribuoj laŭeble estu konataj kaj sub H_0 kaj sub H_1. En nia ekzemplo, taŭga test-adedo estas la variablo k. Se H_0 veras, tiam $k \sim BIN(25, 0.45)$. Ĉiuj entjeroj inter 0 kaj 25 povas realiĝi, sed iliaj probabloj estas tre malsimilaj, kiel oni povas vidi en la malsupra tabelo kaj grafike prezentite en figuro 7.

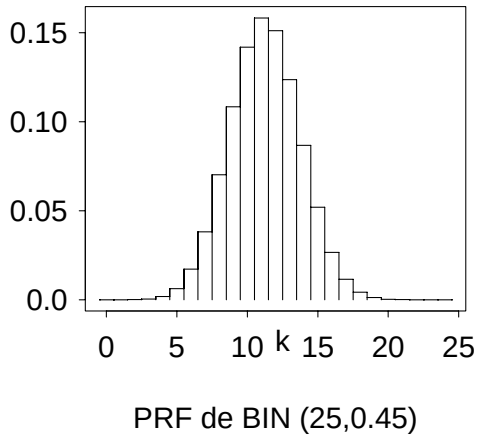
Probablo-funkciono de $BIN(25, 0.45)$

k	< 4	4	5	6	7	8	9	10	11	12	13
PRF	0.00	0.00	0.01	0.02	0.04	0.07	0.11	0.14	0.16	0.15	0.12

k	14	15	16	17	18	> 18
PRF	0.09	0.05	0.03	0.01	0.00	0.00

Plej probablaj estas la realigaĵoj 10, 11, 12, dum ke la probabloj por 0, 1, 2 kaj 20 ... 25 estas tiom malgrandaj, ke oni ne plu povas prezenti ilin sur la grafikaĵo (< 0.0002). Tamen ankaŭ tiuj valoroj ne ekzakte egalas nulon, do, eĉ se H_0 veras, ili povas, kvankam kun malgrandega probablo, realiĝi. Se ni decidus malakcepti la testhipotezon nur tiam, kiam la esplor-rezulto tute ne estas akordigebla kun ĝi, tiam ekestus la dilemo, ke ni povus malakcepti la testhipotezon **neniam**, tutegale kian rezulton ni ricevus. Ĉe tia supersingardema sinteno ne eblas efektivigi statistikan teston, estus do eĉ superflue efektivigi iun esploron.

FIGURO 7



Evidente tia rigida sinteno ne estus prudenta. Ni devas rezigne kontentiĝi pri tio, ke 100%-a certeco ne estas ebla, sed ke ni povas akiri nur iun gradon de probablo. La statistiko ebligas al ni apriore fiksi la probablon, je kiu minimume estu vere, kion ni konkludas el la rezultoj de niaj esploroj. Kutime oni postulas certecon de 95%, oni do vole-nevole toleras 5%-an probablon de eraro, pli ekzakte 5%-an **probablon de erara malakcepto** de la testhipotezo, do de erara akcepto de la alternativ-hipotezo en kazoj, kie la testhipotezo estas vera. Tiun erarprobablon (supre egaligita al 0.05) oni indikas per la simbolo α.

Ni estas nun konstruontaj nian teston tiamaniere, ke ĝi plenumu la postulon de maksimume 5%-a probablo de erara malakcepto. Por tio ni kalkulas la **malakceptigan adedon** (m.a.a). Ĝi estas la probablo sub H_0 por tio, ke oni ricevas test-adedon, kiu estas same aŭ eĉ pli malprobabla kiom la fakte observita, kaj kiu samtempe estus pli

probabla sub H_1 (kondiĉe ke la testrezulto ne estas eĉ malpli akordigebla kun H_1 ol kun H_0, vidu malsupre la situacio ĉe "unuflanka alternativ-hipotezo"). Ĉar la m.a.a. estas la sumo de la probabloj en la vostoj de la distribuo-funkciono, oni nomas ĝin ankaŭ vosta probablo. Se en nia ekzemplo ni observintus $k = 6$, tiam la m.a.a. estus $P(k \leq 6$ aŭ $k \geq 17 = P(0,..,6) + P(17,..,25) = 0.03 + 0.01 = 0.04$. Se la m.a.a. malpliedas* al α, tiam oni malakceptas la testhipotezon kaj akceptas anstataŭe la alternativ-hipotezon. La aro de eblaj realigaĵoj, kies m.a.a $\leq \alpha$, nomiĝas **malakceptiga regiono** (m.a.r.), ties komplemento **ne-malakceptiga regiono**. Oni do malakceptas la testhipotezon, se kaj nur se la test-adedo situas en la malakceptiga regiono. Ĉe nia ekzemplo la m.a.r. konsistas el la du intervaloj [0,6] kaj [17,25]. Tiu procedmaniaro certigas, ke la probablo de erara malakcepto malpliedas ol α, ĉar ni povas aserti: se la testhipotezo estas vera, (se la substanco ne influas la ĝermivon kaj sekve $p = 0.45$), tiam la test-adedo k situos tre verŝajne en la regiono super 6 kaj sub 17 - ne nur tre verŝajne, sed kun probablo plieda al 0.95. Sekve la probablo, ke $k \leq 6$ aŭ $k \geq 17$ kaj ni tial erare asertas ke $p \neq 0.45$, malpliedas al 0.05. Kontraŭe, se la alternativhipotezo veras (se la substanco reduktas aŭ pligrandigas la ĝermivon kaj konsekvence la ĝerma probablo ne egalas al 0.45), tiam la probablo por ($k \leq 6$ aŭ $k \geq 17$) estas pli granda.

Testrezulto, kiu malakceptigas la testhipotezon, nomatas **signifika** (signifika laŭ PIV: "tia, ke ĝi ne dependas de hazardo aŭ neekzakteco de nombroj"). La testrezulton oni nomas "simple", "tre" resp. "treege" signifika, se la koncerna α egalas al 0.05, 0.01 resp. 0.001.

La ĝis nun pritraktita ekzemplo havas hipotezo-paron, kies H_1 estas **duflanka** (ĉar ĝi postulas: $p < 0.45$ **aŭ** $p > 0.45$), kaj kies malakceptiga regiono tial konsistas el du partoj. Sed ofte oni havas **unuflankan** alternativ-hipotezon, ekz.: $p < 0.45$ (la substanco X reduktas la ĝermivon). Tiam la testhipotezo estas aŭ (ebleco A): $p = 0.45$ aŭ (ebleco B): $p \geq 0.45$. Eblecon A oni elektas, se oni scias, ke p neni-okaze povas plii ol 0.45; eblecon B - se la alternativo $p > 0.45$ ne interesas.

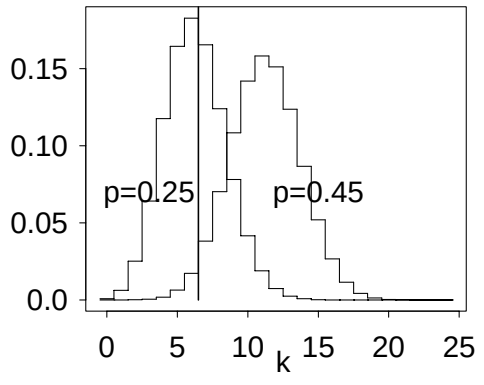
Konsiderante eblecon A, estas evidente, kiamaniere kalkuli la malakceptigan adedon kaj sekve konstrui la malakceptigan regionon: Sub H_0 validas, ke $k \sim BIN(25, 0.45)$. Se k malplias ol sia H_0-a ekspekto, do se $k < 25 * 0.45$, tiam m.a.a.(k) = P(0,..,k). Oni ĉi-foje **ne** rajtas adicii la probablojn de realigaĵoj pli grandaj ol $25 * 0.45$, ĉar tiuj ja estas malprobablaj sub H_0, sed ankoraŭ pli malprobablaj sub nia unuflanka H_1. Se oni ekz. observintus $k = 6$, tiam la m.a.a. estus ĉi-foje egala al $P(k \leq 6) = 0.03$. Se k plias ol sia ekspekto (ekz. se $k = 20$), tiam la rezulto kontraŭas al H_1 eĉ pli ol al H_0, kaj la m.a.a ne plu kalkuleblas (vidu supre la interkrampan rimarkigon ĉe la difino de m.a.a. !). La m.a.r. de la unuflanka $H_1 : p < 0.045$ estas do la intervalo [0,7] , se $\alpha = 0.05$. Tiu procedmaniero validas ankaŭ por ebleco B ($H_0 : p \geq 0.45$). Tiu H_0 estas kunigaĵo de nefinie multaj simplaj testhipotezoj: $H_0 = \cup_{p^\sim \in [0.45,1]} p = p^\sim$, el kiuj la hipotezo $H_0^* : p = 0.45$ estas la plej proksima al H_1. Se la erarprobablo α validas por tiu H_0^*, tiam la erarprobabloj por **ĉiuj** komponantoj de H_0 malpliedas ol α, ĉar $P(k \leq 7$ se $p > 0.45) < P(k \leq 7$ se $p = 0.45)$.

Ni povas resumi: Por ĉiaj hipotezoparoj (duflankaj kaj unuflankaj laŭ ebleco A aŭ B) validas: Se la m.a.a. de la test-adedo estas difinita kaj malplieda ol α (aŭ, samsignife, se la test-adedo situas en la α-rilata malakceptiga regiono), tiam la testo

liverintas signifikan rezulton je la signifo-nivelo α kaj ni rajtas malakcepti la testhipotezon, ni statistike pruvis (je difinita erarprobablo) la alternativ-hipotezon. Se la test-adedo situas en la ne-malakceptiga regiono, tiam ni ne povas malakcepti la testhipotezon, ŝajnaj kontraŭdiroj inter ĝi kaj la testrezulto povas esti kaŭzitaj de hazardaj fluktuaĵoj.

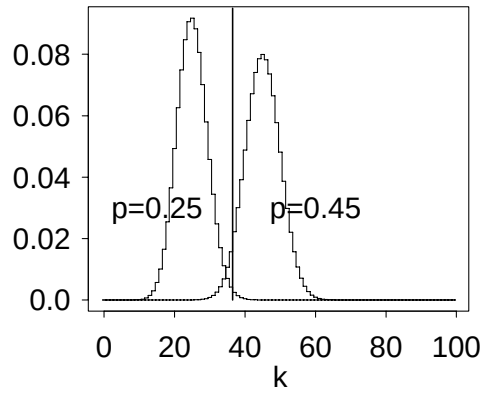
El tio kelkaj homoj konkludas, ke nesignifika testrezulto pruvas la testhipotezon. Tiu konkludo estas tute erara. Por detaligi tion, ni nun analizu (helpe de la hipotezoparo : $H_0 : p = 0.45$; $H_1 : p < 0.45$), kiamaniere la testo reagas, se la **alternativhipotezo** veras. Ni supozu, ke la vera valoro de la parametro p ne estu 0.45, kiel supozas H_0, sed 0.25, do jam konsiderinde malpli granda. En figuro 8 estas grafike prezentata la probablo-funkciono $BIN(25, 0.25)$ apud tiu de $BIN(25, 0.45)$. Oni vidas, ke nun jam estas sufiĉe granda la probablo, ke la test-adedo situas en la m.a.r. inter 0 kaj 6 (tiu probablo estas 0.56) - sed ankaŭ nun kun probablo de 0.44 la test-adedo tamen situas en la nemalakceptiga regiono. De tio ni devas konkludi, ke eĉ tiom granda devio de la testhipotezo kiom la diferenco inter 0.25 kaj 0.45 restas ne-detektota kun probablo de 0.44. Klare videblas, ke H_0 tute ne estas pruvita per tio, ke la test-adedo situas en la m.a.r. Tial oni hodiaŭ ne plu ŝatas nomi tiun regionon simple akceptiga regiono, kiel oni faris fakte antaŭe, ĉar tiu termino estas tro erariga.

FIGURO 8



PRF de BIN (25,0.45) kaj BIN (25,0.25)

FIGURO 9



PRF de BIN (100,0.45) kaj BIN (100,0.25)

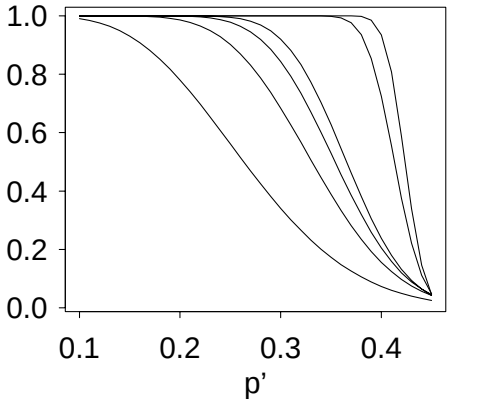
Se oni volas plialtigi la probablon por detekti deviojn de la testhipotezo sen samtempe pligrandigi la probablon **erare** malakcepti ĝin, oni devas pligrandigi la amplekson de la samplo. Figuro 9 montras la probablo-funkcionojn de $BIN(100, 0.45)$ kaj $BIN(100, 0.25)$. Kun $n = 100$ la m.a.r. etendiĝas de 0 ĝis 36. Se la vera parametro p egalas al 0.25, tiam la test-adedo k kun tre granda probablo, nome egala al 0.005, situas en la malakceptiga regiono. Se oni ricevas nesignifikan testrezulton, oni tial povas prave aserti, ke p kun

granda probablo ne estas egala al 0.25 (aŭ eĉ pli malgranda). Sed ankaŭ nun oni ne povas aserti, ke oni pruvis la testhipotezon, ĉar se la parametro p egalus 0.40, do estus pli proksime al la laŭtesthipoteza 0.45, tiam kun probablo de 0.76 la test-adedo situus en la ne-malakceptiga regiono, kaj la rezulto estus ne-signifika.

Kun ĉiu testo estas ligitaj du specoj de eblaj eraroj. Ĝis nun ni estas priparolintaj la eraran malakcepton de vera testhipotezo. Ĝin oni nomas **eraro de unua speco**, ĝia probablo α estas fiksita de la esploranto. Sed, kiel ni ĵus vidis, ekzistas ankaŭ la eblo de erara ne-malakcepto de malvera testhipotezo. Ĝin oni nomas **eraro de dua speco**, ĝian probablon oni indikas per β. β dependas de la testparametroj (en nia ekzemplo de la parametro n), de la unua-speca erarprobablo kaj de la vera grando de la testita parametro.

Oni povas difini funkcionon $DF(p')$, kiu indikas por donita argumento p', kun kiom granda probablo la test-adedo k situas en la m.a.r., tiel liverante signifikan rezulton, se $p = p' : DF(p') = P(k \in \text{m.a.r.}) = 1-\beta$ (donitaj $[H_0, H_1]$, n, α, p'). La funkciono do deskriptas la ivon de la testo detekti erarajn testhipotezojn, tial ni nomu ĝin **detektiva funkciono**. Figuro 10 montras grafike la detektivajn funkcionojn por $[H_0 : p = 0.45\ ; H_1 : p < 0.45], \alpha = 0.05$ kaj diversaj valoroj de n (de maldekstre dekstren n = 25, 50, 75, 100, 500, 1000).

FIGURO 10



detektivaj funkcionoj por diversaj n

IV.c.2 Signuma testo

Helpe de la dunomiala distribuo oni povas bone kompari du similsignifajn atributojn ĉe la sama esplorunuo.

Ekzemplo : Oni volas scii, ĉu du genotipoj de piceo malsamas laŭ la konkurivo dum la plej frua evoluo. Tion oni povas esplori per jena eksperimento: En n potoj oni ĝermigas po unu pice-semon de ambaŭ genotipoj, kaj post difinita tempo, kiam la konkuro jam influintas la evoluon, oni determinas por ĉiu planto la sukceson en la konkur-batalo per atributo skaligita almenaŭ ordinale, ekzemple per la grando. Estu k_1 resp. k_2 la nombro de potoj, en kiuj la planto de genotipo 1 resp. 2 estis la pli sukcesa (en $n - k_1 - k_2$ potoj ne rimarkeblis diferenco inter la plantoj). La test-adedo $k_1 \sim BIN(k_1 + k_2, p)$. Por esplori, ĉu unu el la du genotipoj estas pli konkuriva ol la alia, oni starigas la hipotezoparon $(H_0 : p = 0.5; H_1 : p \neq 0.5)$, kiu testeblas kiel estas

supre deskriptite. La nomo de ĉi-tiu testo estas **signumo-testo**, ĉar ĝi uzas nur la signumojn de la diferencoj inter la atributoj.

IV.c.3 t-testoj

La sekva testo pritraktas atributon y, kiu estas (almenaŭ proksimume) gaŭse distribuita kaj ne havas distantojn : $y \sim N(\mu, \sigma^2)$. La atributo y povus esti alto aŭ diametro de samaĝaj ekzempleroj de iu planto-specio, fiziologia adedo ĉe medicina esploro aŭ simila. Unue ni volas testi la hipotezon $\mu = \mu_0$, kie μ_0 estas nombro donita de la esploranto.

Se $(y_1, ..., y_n)$ estas aleatora samplo (samplo rilata al aleatora specimeno) el $N(\mu, \sigma^2)$, tiam la grando $T_1 = (\bar{y} - \mu) * \sqrt{n}/s_y$, kie s_y estas la sampla ordinara devio de y, sekvas tiel-nomatan **t-distribuon** $t(l.g. = n - 1)$. La t-distribuo havas unu parametron, nomatan **libero-grado** $l.g.$, kiu en praktikaj aplikoj estas pozitiva entjero kaj en ĉi-tiu kazo egalas al $n - 1$. La t-distribuo estas simetria ĉirkaŭ 0; ju pli granda estas ĝia libero-grado, des pli ĝi similas al $N(0, 1)$. Ĝiaj pluriloj estas troveblaj en tabeloj kaj facile kalkuleblaj per komputiloj. Jena tabelo montras la 95%- kaj la 97.5%-plurilojn por kelkaj libero-gradoj

l.g. =	1	2	3	4	5	6	7	8	9	10	11	12
γ=0.950	6.31	2.92	2.35	2.13	2.02	1.94	1.90	1.86	1.83	1.81	1.80	1.78
γ=0.975	12.71	4.30	3.18	2.78	2.57	2.45	2.37	2.31	2.26	2.23	2.20	2.18

l.g. =	13	14	15	16	17	18	19	20	30	40	60	∞
γ=0.950	1.77	1.76	1.75	1.75	1.74	1.73	1.73	1.73	1.70	1.68	1.67	1.65
γ=0.975	2.16	2.15	2.13	2.12	2.11	2.10	2.09	2.09	2.04	2.02	2.00	1.96

Kiel supre montrite ĉe $N(0, 1)$, la plurilojn por $\gamma < 0.5$ oni kalkulas per
$PLF[t(l.g.)](\gamma) = -PLF[t(l.g)](1 - \gamma)$, ekz.
$PLF[t(10)](0.025) = -PLF[t(10)](0.975) = -2.23$.

Por testi $H_0 : \mu = \mu_0$, oni kalkulas la test-adedon $T_1(\mu_0) = (\bar{y} - \mu_0) * \sqrt{n}/s_y$. Se H_0 veras, do se μ_0 fakte estas la ekspekto de y, tiam la test-adedo estas distribuita laŭ $t(n - 1)$. Se $\mu < \mu_0$, tiam $T_1(\mu_0)$ havas la tendencon esti malpli granda ol sub H_0; se $\mu > \mu_0$ - esti pli granda. La m.a.a. kaj la m.a.r. por $T_1(\mu_0)$ kalkuliĝas jene: Se $H_1 : \mu < \mu_0$, tiam la m.a.r. estas la intervalo $[-\infty, PLF[t(n - 1)](\alpha)]$, ĉar en ĉi-tiu malsupra parto de la reela akso situas la valoroj, kiuj estas malprobablaj sub H_0 kaj samtempe pli probablaj sub H_1. La m.a.a. por negativa test-adedo $T_1(\mu_0)$ estas $DIF[t(n-1)](T_1)$, por pozitiva T_1 m.a.a. ne estas difinita. Ekzemplo : Por samplo kun n=15, la m.a.r. por α=0.05 estas la intervalo $[-\infty, -1.76]$, kaj se $\bar{y}$=9.60 kaj s_y=0.77, tiam por μ_0=10 ni ricevas: $T_1(10) = (-0.40 * \sqrt{15})/0.77 = -2.01$. El la tabelo oni vidas, ke $0.025 < m.a.a < 0.050$ (do simple signifika), ĉar $-2.15 < 2.01 < 1.76$. La ekzakta valoro por m.a.a., kalkulita per komputilo, estas 0.032. Inverse estas, se $H_1 : \mu > \mu_0$. Tiam la m.a.r. estas la intervalo $[PLF[t(n-1)](1-\alpha), \infty]$, por n=15 sekve $[1.76, \infty]$. La m.a.a. por pozitiva test-adedo $T_1(\mu_o)$ estas $1 - DIF[t(n - 1)](T_1(\mu_0))$, por negativa T_1 m.a.a. ne estas difinita. Se $H_1 : \mu \neq \mu_0$ (duflanka alternativo), tiam la m.a.r. konsistas el la du intervaloj $[-\infty, PLF[t(n-1)](\alpha/2)]$ kaj $[PLF[t(n-1)](1-\alpha/2), \infty]$, ekz. por n=15 kaj α=0.05 el $[-\infty, -2.15]$ kaj $[2.15, \infty]$. La m.a.a $(T_1(\mu_0)) =$
$DIF[t(n-1)](-|T_1(\mu_0)|)+(1-DIF[t(n-1)](+|T_1(\mu_0)|)) = 2*DIF[t(n-1)](-|T_1(\mu_0)|)$.

Per la sama testo oni povas kompari ĉe la sama esplorunuo du similsignifajn atributojn, kiuj estas proksimume gaŭse distribuitaj kaj ne havas distantojn. La celvariablo $d = y_2 - y_1$ estas ĉi-kaze distribuita laŭ $N(\mu_d, \sigma_d^2)$. La komparon oni performas per hipotezoparo kun $H_0 : \mu_d = 0$. Se la premisoj por ĉi-tiu testo estas plenumitaj, do se almenaŭ prosimume d estas gaŭse distribuita kaj forestas distantoj, tiam la t-testo estas pli detektiva ol la signuma testo, t.e. ĉe malveraj testhipotezoj ĝi pli ofte liveras signifikajn rezultojn. Sed se en la populacio povas ekzisti distantoj kaj/aŭ la atributoj estas oblikve distribuitaj, tiam la rezultoj de la t-testo ne estas fidindaj, ĉi-kaze oni preferu la signuman teston.

Ofte oni volas kompari la parametrojn de du sub-populacioj. Imagu problemon similan, sed ne egalan al la ĵus pritraktita ekzemplo: Oni volas kompari la kreskorapidon, ne influitan de konkurenco, de du genotipoj de piceo. Por tion pritrakti oni ĝermigas po unu semon de genotipo 1 en n_1 potoj kaj de genotipo 2 en n_2 potoj (kutime, sed ne ĉiam, $n_1 = n_2$). Post kelka tempo oni mezuras taŭgan atributon y ĉe ĉiuj plantoj. Estu konate, ke almenaŭ proksimume $y_1 \sim N(\mu_1, \sigma^2)$ kaj $y_2 \sim N(\mu_2 = \mu_1 + d, \sigma^2)$. Tio implicas, ke ĉiuj semoj ĝermis (alie la distribuo de la atributo estus oblikva!), ke ne ekzistas distantoj, kaj ke y_1 kaj y_2 havas la saman variancon. Se tiuj kondiĉoj estas plenumitaj, tiam la grando

$$T_2 = (\bar{y}_1 - \bar{y}_2 - d) * \sqrt{n_1 + n_2 - 2} / \sqrt{1/n_1 + 1/n_2} * \sqrt{SQy_1 + SQy_2)} \sim t(n_1 + n_2 - 2)$$

Por testi la hipotezon : $d = d_0$, oni kalkulas la test-adedon $T_2(d_0)$, anstataŭigante en la supra difino d per d_0. Se H_0 veras, tiam $T_2(d_0) \sim t(n_1 + n_2 - 2)$, kaj la kalkulado de la m.m.a. kaj de la m.a.r. performatas kiel estas deskriptite en la antaŭa ekzemplo. La plej ofta apliko de tiu-ĉi testo estas la ekzamenado, ĉu $d = 0$, kio signifus, ke la du subpopulacioj havas la **saman** ekspekton. Kiamaniere oni povas procedi, se la supre indikitaj test-premisoj ne estas plenumataj, ni bedaŭrinde ne povas priparoli ĉi-tie.

Tria grava ekzemplo por la apliko de la t-distribuo estas la testo de hipotezo pri la parametro b_1 ĉe la simpla (t.e. ne-multobla) regresi-analizo, do por testi $H_0 : b_1 = b_{1;0}$ (kutime $b_{1;0} = 0$), kie b_1 estas la parametro de la populacio. Se la dependa variablo y_i estas gaŭse distribuita ĉirkaŭ sia ekspekto $b_0 + b_1 * x_i$ kaj havas variancon konstantan, ne de x_i dependan, tiam $T_b = (\hat{b}_1 - b_1) * \sqrt{SQx * (n-2)} / \sqrt{SQresto} \sim t(n-2)$, kie $SQresto = SQy - \hat{b}_1 * SPxy$. Se H_0 veras, tiam la test-adedo $T_b(b_{1;0}) \sim t(n-2)$, la kalkulado de m.a.a. kaj m.a.r. sekvas denove la saman skemon.

IV.d Konfidenc-intervaloj

Helpe de la ĵus pritraktitaj t-distribuitaj valoroj T_1, T_2, T_b oni povas konstrui **konfidenc-intervalojn** por la respektivaj populaciaj parametroj. Tiuj estas intervaloj, kiuj kun probablo de $1-\alpha$ enhavas la nekonatan populacian parametron. La procedmaniero estu montrita per la ekzemplo de duflanka konfidencintervalo de la ekspekto μ de gaŭse distribuita variablo helpe de T_1. El la fakto, ke $T_1 = (\bar{y} - \mu) * \sqrt{n}/s_y \sim t(n-1)$ ni

povas konkludi, ke kun probablo α validas :

$$\begin{array}{lrcl}
 & PLF[t(n-1)](\alpha/2) & < (\bar{y}-\mu)*\sqrt{n}/s_y & < PLF[t(n-1)](1-\alpha/2) \\
\Leftrightarrow & PLF[t(n-1)](\alpha/2)*s_y/\sqrt{n} & < (\bar{y}-\mu) & < PLF[t(n-1)](1-\alpha/2)*s_y/\sqrt{n} \\
\Leftrightarrow & -\bar{y}+PLF[t(n-1)](\alpha/2)*s_y/\sqrt{n} & < -\mu & < -\bar{y}+PLF[t(n-1)](1-\alpha/2)*s_y/\sqrt{n} \\
\Leftrightarrow & \bar{y}-PLF[t(n-1)](\alpha/2)*s_y/\sqrt{n} & > \mu & > \bar{y}-PLF[t(n-1)](1-\alpha/2)*s_y/\sqrt{n} \\
\Leftrightarrow & \bar{y}-PLF[t(n-1)](1-\alpha/2)*s_y/\sqrt{n} & < \mu & < \bar{y}+PLF[t(n-1)](1-\alpha/2)*s_y/\sqrt{n}
\end{array}$$

Ekzemplo: por samplo-amplekso n=15, $\bar{y}$=9.60 kaj sampla o.d. s_y=0.77 ni ricevas

$$\begin{array}{lrcl}
9.60-PLF[t(14)](1-\alpha/2)*0.77/\sqrt{15} & < \mu & < & 9.60+PLF[t(14)](1-\alpha/2)*0.77/\sqrt{15} \\
\Leftrightarrow \text{ (se } \alpha\text{=0.05 kaj sekve } PLF=2.15) & & & \\
9.60-0.43 & < \mu & < & 9.60+0.43 \\
\Leftrightarrow 9.17 & < \mu & < & 10.03
\end{array}$$

Tio signifas, ke supozeble la vera populacia parametro μ situas inter 9.17 kaj 10.03.

Unuflankan (supre malferman) konfidencintervalon oni konstruas elirante de la malegalaĵo $PLF[t(n-1)](\alpha) > (\bar{y}-\mu)*\sqrt{n}/s_y$
El tio oni ricevas : $\mu > \bar{y}-PLF[t(n-1)](\alpha)*s_y/\sqrt{n}$.

La konfidencintervalojn por d kaj b_1 oni akiras tute analogie helpe de T_2 kaj T_b.

Glosoj por esprimoj ne klarigitaj en la teksto

deskripti : prezenti iun fakton tiamaniere, ke la adresito ĝin komprenu (Propono de Wells). Laŭ PIV tio estus "priskribi", sed tiu vorto estas erariga.

fitotrono : ĉambro por kreskigi plantojn en ekzaktaj medi-kondiĉoj

funkciono : 3-a signifo de "funkcio" en PIV (bildigo en aron de nombroj) (propono fare de mi)

(mal)pliedi : a (mal)pliedas ol b signifas, ke $a \leq b$ (resp. $a \geq b$) (propono de C.O.Kiselman)

Adreso de la aŭtoro :
OProf. H.D. Quednau dr., Forstwiss. Fakultät der TU,
Am Hochanger 13, D-85354 Freising
email : quednau@lrz.tu-muenchen.de